

AD-A249 646



AD _____

2

MODELING FOR HUMAN PERFORMANCE ASSESSMENT

FINAL REPORT

R. R. STANNY

NOVEMBER 20, 1991

Supported by

U.S. ARMY MEDICAL RESEARCH AND DEVELOPMENT COMMAND
Fort Detrick, Frederick, Maryland 21702-5012

MIPR 90MM0523

Naval Aerospace Medical Research Laboratory
Naval Air Station
Pensacola, Florida 32508-5700

Approved for public release; distribution unlimited.

The findings in this report are not to be construed as an
official Department of the Army position unless so designated
by other authorized documents

92-11997



92 5 01 87

REPORT DOCUMENTATION PAGE

Public reporting burden for this report is estimated to be 1 hour per report, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the report form. Send comments regarding this burden estimate or any other aspect of this report form, including suggestions for reducing the burden, to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Project Director (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (leave blank)

2. REPORT DATE

20 November 1991

3. REPORT TYPE

Final 1 Nov 89 - 21 Nov 91

4. TITLE AND SUBTITLE

Modeling for Human Performance Assessment

MIPR 90MM0523

6. AUTHOR(S)

R.R. Stanny

63002A

3M263002D995 BG

DA346133

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Aerospace Medical Research Laboratory
Bldg. 1953, Naval Air Station
Pensacola, FL 32508-5700

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Medical R&D Command
Fort Detrick
Frederick, Maryland 21702-5012

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words) This final report describes research performed during Fiscal Years 1990 and 1991 at the Naval Aerospace Medical Research Laboratory. One research line involved developing a generic model of human performance tests, such as those in the Unified Tri-service Cognitive Performance Assessment Battery. Several performance-test models were developed using the plan of the generic task. A second research line focused on a risk identification study of 31 Navy and Marine combat occupations. The purpose was to examine whether knowledge of a stressor's effects on abilities might be used to predict the combat jobs most likely to be affected by the stressor. The results identified certain perceptual-motor, and cognitive abilities that may vary substantially in importance among the occupations. Examples included far vision, spatial orientation, flexibility of closure, rate control and several others. Analyses of stressors for their effects on these abilities may suggest ways to optimize the use of resources by distributing backup personnel, countermeasures, and other risk-mitigating factors among jobs in part according to relative threat magnitudes. A third line of research focused on the development of a MicroSAINT model of an aircraft carrier landing. Although this line of research was interrupted a year early, due to reductions in research funds, a preliminary model was developed and is described in this report. A fourth line of research focused on issues in laboratory-test design and analysis. Studies in this line included (1) an examination of factors contributing to the potential sensitivities of tests to the effects of environmental stressors and (2) an examination of the use of iterated bootstrap resampling in the calculation of Monte Carlo estimates of the precision and significance levels of psychological tests.

Modeling, Human Performance ; RA 5

Unclassified

Unclassified

Unclassified

Unlimited

GENERAL INSTRUCTIONS FOR COMPLETION

The Report Documentation Form (RDM) is used by authors to provide information that this information is consistent with the rest of the report. Additionally, the RDM contains the instructions for filling in each block. The instructions follow. It is important to read within the first few pages of the report the *optical scanning requirements*.

Block 1: Agency Use only (to be filled in)

Block 2: Report Date (to be filled in) (date including day, month, and year available to 1 Jan 88). Must state after 1 Jan 88.

Block 3: Type of Report and Date Received. State whether report is internal or external. If applicable, enter for internal report (1 Jan 88 to Jun 87) or for external report (1 Jul 87 to Jun 88).

Block 4: Title and Subtitle. Enter title from the part of the report that provides the most meaningful and complete description. When a report is prepared in more than one volume, repeat the primary title in the first volume and include subtitle for the other volumes. On classified documents, add the classification in parentheses.

Block 5: Funding Numbers. Leave blank and grant numbers, may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following table:

C	Contract	NA	Agency
G	Grant	NA	Task
PE	Program Element	NA	Agency

Block 6: Author. Enter the name of the person responsible for writing the report. Do not enter the name of the person who prepared the report. Do not enter the name of the person who prepared the report.

Block 7: Address. Enter the address of the person responsible for writing the report.

Block 8: Address. Enter the address of the person responsible for writing the report.

Block 9:

Block 10:

Block 11:

Block 12a: Distribution Code. Enter the distribution code for the report. The distribution code is a three-digit number that indicates the distribution status of the report. The distribution code is entered in the first three digits of the report number.

DOD: Leave blank. Do not enter a distribution code for DOD reports.

DDI: Leave blank. Do not enter a distribution code for DDI reports.

NASA: Leave blank. Do not enter a distribution code for NASA reports.

NHS: Leave blank. Do not enter a distribution code for NHS reports.

Block 12b: Distribution Code

DOD: Leave blank. Do not enter a distribution code for DOD reports.

DDI: Leave blank. Do not enter a distribution code for DDI reports. Do not enter a distribution code for DDI reports.

NASA: Leave blank. Do not enter a distribution code for NASA reports.

NHS: Leave blank. Do not enter a distribution code for NHS reports.

Block 14: Abstract. Include a brief (maximum 200 words) factual summary of the most significant information contained in the report.

Block 15: Subject Terms. Enter the subject terms for the report. Do not enter the subject terms for the report.

Block 16: Subject Terms. Enter the subject terms for the report. Do not enter the subject terms for the report.

Block 17: Subject Terms. Enter the subject terms for the report. Do not enter the subject terms for the report.

Block 18: Subject Terms. Enter the subject terms for the report. Do not enter the subject terms for the report.

Block 19: Subject Terms. Enter the subject terms for the report. Do not enter the subject terms for the report.

Block 20: Subject Terms. Enter the subject terms for the report. Do not enter the subject terms for the report.

LIST OF INDIVIDUALS RECEIVING PAY FROM MIPR NO. 90MM0523

Stanny, R.
 Shamma, S.
 Raynsford, K.
 LaCour, S.
 Travis, K.



Accession For	
NTIS ORCAI	<input checked="" type="checkbox"/>
ERIC FAR	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

FOREWORD

Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

✓ Where copyrighted material is quoted, permission has been obtained to use such material.

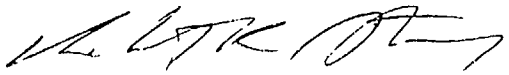
✓ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

✓ Citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement or approval of the products or services of these organizations.

✓ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Animal Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

✓ For the protection of human subjects, the investigator(s) have adhered to policies of applicable Federal Law 45CFR46.

✓ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institute of Health.

 11/7/91

Principal Investigator's Signature Date

CONTENTS

	<u>Page</u>
FOREWORD	iii
1. INTRODUCTION	1
2. RISK ANALYSIS OF 31 NAVAL AND MARINE COMBAT OCCUPATIONS	1
Methods	1
Results and Discussion	2
Conclusions	8
3. SENSITIVITY OF TESTS	9
Discussion	9
Conclusions	14
4. CARRIER LANDING MODEL	16
Model Description	16
5. BOOTSTRAP CONFIDENCE INTERVALS	24
Method	27
Results and Discussion	28
Conclusions	29
REFERENCES	30
BIBLIOGRAPHY	33

1. INTRODUCTION

This Final Report covers work performed at the Naval Aerospace Medical Research Laboratory during Fiscal Years 1990 and 1991. The 1990 work included (a) developing models of human performance tests drawn from the Unified Tri-services Cognitive Assessment Battery and (b) a risk-identification study of 31 naval and marine combat occupations. The 1991 work included (a) a methodological study of test sensitivity, (b) development of a MicroSAINT model of an aircraft carrier landing and (c) a Monte Carlo study of bootstrap data-resampling. The FY 1990 modeling effort is described in detail in Stanny and Shamma (1990), a copy of which is attached to this Report. The FY 1990 study of combat occupations is described in Section 2 of the text. The FY 1991 sensitivity study is described in Section 3 of this report. The carrier-landing study is described in Section 4 of the text. The Monte Carlo study of bootstrap resampling is described in Section 5.

2. RISK ANALYSIS OF 31 NAVAL AND MARINE COMBAT OCCUPATIONS

Resources at hand are never sufficient to ensure that all combat personnel will be protected from all possible hazards. Thus, strategies must be developed to estimate the proportion of resources that should be devoted to countermeasures and to allocate those resources as well as possible. A basic problem in the development of such strategies is to identify those personnel most threatened by each different potential hazard. The assumption that different stressors will affect different abilities suggests that a taxonomy based on skills and abilities would be valuable in this regard (Cooper, Schemmer, Fleishman, Yarkin-Levin, Harding, & McNelis, 1987).

Given that the impact of a specific stressor can be expressed as a pattern of changes in a set of abilities, it should be possible to derive the relative impact of a stressor on each of a given set of jobs. The analyses presented here are based on the assumption that the magnitude of a stressor's threat to performance on task i increases with the number and importance of the skills affected by the stressor. That is,

$$t_i = f_k(\sum_j e_{jk} s_{ij}), \quad (1)$$

where t_i represents the threat to the i th task, f_k is a monotonically increasing function that may differ among stressors, e_{jk} is a dummy variable equal to 1 if stressor k affects skill j and 0 otherwise, and s_{ij} is the importance of the j th skill to task i . I do not mean to imply that Equation 1 should be regarded as a general model of stressor effects. It is, however, an assumption that should be consistent with a range of such models.

In this report, I will discuss three exploratory analyses of a data base of Navy and Marine air combat occupations (Cooper et al., 1987). Each was performed with an eye to determining which skills might be most informative in predicting differential risks posed by environmental stressors. The first analysis described here comprised an examination of the variation in the skills' importance ratings across jobs. The second analysis was performed by identifying clusters of jobs related by similarities in their patterns of skill requirements and then determining the variables that best distinguished between the job clusters. The intuition motivating this analysis was that using clusters of similar jobs as units of analysis might yield more stable predictions than those derived from analyses of individual jobs. In the third analysis, I identified clusters of skills related by their patterns of association across jobs and then examined the degree to which these skill clusters distinguished between the groups of jobs previously identified.

Methods

Task analysis data. The data base of occupational task analyses used in the present study was developed by Cooper et al. (1987). The data base contains task analyses of 31 naval and marine combat jobs. The

information on each job includes a rating of the importance of each of 44 skills and abilities to the performance of each job. These ratings were developed through interviews with experienced job incumbents.¹ The rating scale ranged from one to seven (least to most important). The list of jobs is given in Table I; the list of skills for which the jobs were rated is in Table II.

(U) Table I. Navy and Marine Aviation Occupations in the Data Base. (U)

Aviation Boatswain's Mate	Helicopter Crew Chief
Aviation Electrician's Mate	Hospital Corpsman
Aviation Structural Mechanic	Landing Signal Officer
Aviation Ordnanceman	Marine Bulk Fuel Operator
Aviation Organizational Maint. Officer	Marine Helicopter Pilot
Aviation Fire Control Technician	Marine Harrier Pilot
Aviation Electronics Technician	Machinist Mate
Bombardier Navigator	Marine Prop Pilot
Catapult & Arresting Gear Officer	Navy Helicopter Pilot
Cryptologic Technician	Radioman
Data Systems Technician	Helicopter Search & Rescue Crew Member
Electrician's Mate	Sonar Technician
Electronics Technician	Torpedoman's Mate
Electronic Warfare Technician	Tactical Pilot
Fire Controlman	SEALS
Gunner's Mate	

Statistics. Principal Components Analyses (PCAs) were performed using BMDP 4M (Factor Analysis; Dixon, Brown, Engelman, Hill, & Jennrich, 1988). Recall that PCA yields a component for each variable. The first principal component (PC) extracted from the correlation matrix corresponds to the linear combination of the original variables that accounts for the largest proportion of the variance in the data. Subsequent components are statistically independent and account for smaller and smaller proportions of the variance. The scree test (the method of rootstaring; Cliff, 1987) was used to identify components that appeared to represent real phenomena. These PCs were then rotated by the varimax procedure. Varimax rotation produces components whose squared correlations with the original variables have the largest possible collective variance. This tends to produce "simple" components, components strongly correlated with a few of the original variables and weakly correlated with the others. Discriminant analyses (DAs) were performed with BMDP 7M (Dixon et al., 1988). Details specific to individual analyses are described in the next section.

Results and Discussion

Table III contains the mean rated importance of each skill, calculated across jobs, in the Cooper et al (1987) data base. The entries in Table I are sorted in descending order of average rating. One should be cautious about interpreting these means as general measures of "importance" because they are strongly influenced by the makeup of the specific sample of jobs selected for inclusion in the data base. The foregoing having been said, the head of the list is dominated by a set of perceptual/cognitive variables. The middle of the list contains a number of variables having to do with coordination, dexterity, and spatial orientation. Strength and stamina

¹The data base also contains information on subtasks of jobs. Only the overall skills-and-abilities ratings were analyzed in this study.

(U) Table II. Skills and Abilities Rated for Each Job. (U)

1 Oral comprehension	23 Time sharing
2 Written comprehension	24 Rate control
3 Oral expression	25 Arm-hand steadiness
4 Written expression	26 Manual dexterity
5 Fluency of ideas	27 Finger dexterity
6 Originality	28 Speed of limb movement
7 Memory	29 Static strength
8 Problem sensitivity	30 Dynamic strength
9 Mathematical reasoning	31 Explosive strength
10 Number facility	32 Trunk strength
11 Logical reasoning	33 Muscular flexibility
12 Information ordering	34 Equilibrium
13 Speed of closure	35 Gross body coordination
14 Flexibility of closure	36 Stamina
15 Spatial orientation	37 Near vision
16 Visualization	38 Far vision
17 Perceptual speed	39 Color vision
18 Control precision	40 Night vision
19 Multi-limb coordination	41 Depth perception
20 Reaction time	42 Glare sensitivity
21 Choice reaction time	43 General hearing
22 Selective attention	44 Sound localization

variables tend to be found in the lower third of the list. Notable exceptions to the preceding generalizations are math and writing, which are rated as comparatively unimportant. Reading, however, is rated as important.

Table IV contains a list of skills sorted in order of decreasing variability (across jobs) of their importance ratings. This list is of particular interest because the accuracy of predicting which jobs are likely to be affected by a stressor should increase with the systematic variance (across jobs) in the importance of the affected skills. This point can be understood by reference to Equation 1. Examining Equation 1, one can see that, with other factors (including measurement error) held constant, the spread in threat magnitude across jobs, $\text{var}(t_i)$, increases with the job-to-job variance in $\text{var}(s_{ij})$, the importance of a threatened skill. As the spread in threat magnitudes increases, for any reason other than an increase in measurement error, the accuracy of predicting those jobs for which the threat exceeds a critical value should also increase.

The most variable skills on the list of Table IV are a set of perceptual, psychomotor, and strength skills. Of note, most of the cognitive skills fall near the bottom of this list. This suggests that, at least in the present sample of jobs, it may prove easier to predict differential threats to performance from effects of stressors on perceptual and strength variables than from effects of stressors on cognitive variables.

I searched for clusters of related jobs by performing a PCA of a correlation matrix whose row and column headings were the 31 combat jobs. Each element, r_{ij} , of this matrix was, thus, the correlation between the 44 skill ratings of jobs i and j . High values of r_{ij} indicated jobs with similar skill requirements. This procedure resembles Q-factor analysis, a technique sometimes used in studies of individual differences (Guilford, 1954). Examining Fig 1., one can see that by the time the seventh PC was extracted, the magnitudes of the eigenvalues had decreased to a value effectively equal to 1.0. This value is $1/31$ of the total variance (the 31

(U) Table III. Mean Skill Ratings Across Jobs. (U)

Skill	M	Skill	M
Selective attention	6.00	Fluency of ideas	4.35
Problem sensitivity	5.94	Finger dexterity	4.03
Near vision	5.87	Color vision	3.97
Time sharing	5.74	Far vision	3.97
Written comprehension	5.35	Speed of closure	3.90
Night vision	5.26	Originality	3.90
Memory	5.23	Number facility	3.90
Reaction time	5.00	Muscular flexibility	3.84
Logical reasoning	4.94	Visualization	3.74
Oral expression	4.94	Glare sensitivity	3.71
Information ordering	4.90	Static strength	3.68
Flexibility of closure	4.84	Rate control	3.45
General hearing	4.84	Sound localization	3.42
Control precision	4.74	Trunk strength	3.26
Depth perception	4.74	Gross body coordination	3.26
Oral comprehension	4.68	Arm-hand steadiness	3.26
Perceptual speed	4.61	Speed of limb movement	3.23
Multi-limb coordination	4.58	Written expression	3.19
Choice reaction time	4.58	Dynamic strength	3.00
Manual dexterity	4.55	Stamina	2.52
Spatial orientation	4.52	Mathematical reasoning	2.10
Equilibrium	4.45	Explosive strength	1.94

variables in the analysis were standardized so that each had unit variance). Because factors with unit eigenvalues account for no more variance than one of the original variables, nothing is to be gained by considering factors beyond the sixth. Indeed, the plot of eigenvalue magnitude versus eigenvalue number seems to contain a break-point in the vicinity of factor 3-5, which suggests that, perhaps, only the first four factors are real (Cliff, 1987).

Table V shows the clusters of jobs that loaded on (correlated in excess of 0.5 with) each of the four PCs. A group of technical jobs are associated with the first PC. An examination of this cluster suggests that the jobs in it are fairly high in their demands for logical analysis. The second cluster is dominated by pilot occupations and some closely related jobs. The third cluster is dominated by mechanical-technical jobs. The SEALS formed their own fourth cluster. Two jobs did not load to the criterion 0.5 on any of the PCs.

I used linear discriminant analysis (DA) to measure the distances between job clusters in terms of various combinations of the skill ratings. This appeared to be the direct approach to identifying variables that might discriminate between the clusters. Furthermore, it was unclear that simply factoring the skills intercorrelation matrix would

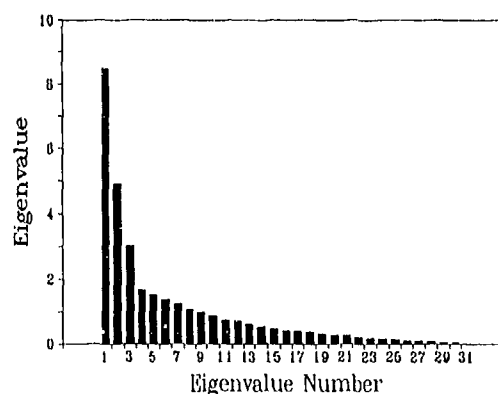


Figure 1. Eigenvalues from the principal components analysis of jobs.

(U) Table IV. Variation in Skill Requirements Across Jobs. (U)

<u>Skill</u>	<u>sd</u>	<u>Skill</u>	<u>sd</u>
Far vision	2.85	Reaction time	1.92
Rate control	2.72	Speed of limb movement	1.90
Stamina	2.66	General hearing	1.82
Glare sensitivity	2.49	Choice reaction time	1.74
Trunk strength	2.46	Finger dexterity	1.69
Depth perception	2.46	Flexibility of closure	1.61
Sound localization	2.42	Mathematical reasoning	1.59
Visualization	2.40	Perceptual speed	1.58
Static strength	2.35	Multi-limb coordination	1.56
Dynamic strength	2.29	Color vision	1.53
Explosive strength	2.29	Manual dexterity	1.52
Arm-hand steadiness	2.24	Written comprehension	1.33
Speed of closure	2.19	Number facility	1.23
Fluency of ideas	2.12	Memory	1.21
Gross body coordination	2.09	Oral comprehension	1.06
Originality	2.08	Information ordering	1.06
Spatial orientation	2.06	Time sharing	0.98
Equilibrium	2.06	Oral expression	0.95
Muscular flexibility	2.05	Near vision	0.83
Night vision	1.95	Selective attention	0.80
Written expression	1.94	Problem sensitivity	0.80
Control precision	1.93	Logical reasoning	0.72

produce similar results. The predictor variables used in the DA were the job skill-requirement levels. The grouping variable was job-cluster membership--the number of the PC with which each job was correlated. Cluster four of the jobs PCA was not used in the DA because it contained only the SEALS. The two uncategorized jobs also were not used. Thus, the resulting prediction equations were linear combinations of the original skill-requirement values that maximized the overall Euclidean distance (in within-group σ units) between the means of the groups defined by the PCA.

I added one variable at a time to the prediction equation by forward stepping. A criterion $F(2,25)$ -to-enter of 9.12 was used to control the entry of variables. This is the Bonferroni-corrected critical value of F that yields an experimentwise significance level of $p \leq .05$ when 44 such F ratios are available for comparison. (Note that, because the job clusters were not determined according to a priori criteria, this significance level may not reflect the actual significance of the DA.)² Five skills produced F ratios greater than 9.12. These were far

²An additional problem is posed by the fact that the number of skills (predictors) in the database exceeds the number of jobs (cases). Hence, the significance of the full-rank discriminant function cannot be calculated. This makes it difficult to assess the significance of the discriminators because the most compelling way to establish that one or more skills significantly distinguishes among the job clusters would be to establish the significance of the full-rank prediction equation(s). (See Larzelere and Muliak, 1977, for a discussion of this issue in the related context of multiple regression). A partial solution is to calculate the significance of prediction equations containing subsets of prespecified size, l , of the original m predictor variables. (Unfortunately, in exploratory analyses one can rarely supply an a priori rationale for setting l to any particular value, with the possible exception of 1.) For a subset of size $l = 1$ selected from m candidate predictors, a conservative, Bonferroni-style significance level can be estimated by determining p in the usual way and multiplying by m . For

(U) Table V. Clusters of Occupations obtained by Principal Components Analysis. (U)

Cluster 1

Aviation Electrician's Mate
Aviation Electronics Technician
Aviation Fire Control Technician
Aviation Organizational Maintenance
Officer
Cryptologic Technician
Data Systems Technician
Electrician's Mate
Electronic Warfare Technician
Electronics Technician
Gunner's Mate
Hospital Corpsman
Radioman
Sonar Technician

Cluster 3

Aviation Boatswain's Mate
Aviation Ordnanceman
Aviation Structural Mechanic
Machinist Mate
Torpedoman's Mate

Cluster 2

Bombardier Navigator
Catapult and Arresting Gear
Officer
Helicopter Crew Chief
Helicopter Search and Rescue
Crew Member
Landing Signal Officer
Marine Harrier Pilot
Marine Helicopter Pilot
Marine Prop Pilot
Navy Helicopter Pilot
Tactical Pilot

Cluster 4

SEALS

vision, spatial orientation, arm-hand steadiness, rate control, and glare sensitivity. Far vision yielded the largest value of $F(2,25) = 36.00$ and was, thus, entered into the prediction equation. Far vision would seem to distinguish flight-related jobs from other occupations. Consistent with this observation, when far vision was entered into the equation, the F ratios for entering spatial orientation, rate control, and glare sensitivity (other clearly flight-related skills) dropped precipitously, from respectable values of 9.37, 10.06, and 10.68 to 0.08, 0.94, and 0.17, respectively.³ The drop suggests that the information they contained was redundant to the prediction equation.

With far vision in the prediction equation, a criterion $F(2,24)$ -to-enter of 9.20 was used to control the entry of further variables. This is the Bonferroni-corrected critical value of F that yields an experimentwise significance level of $p \leq .05$ when 43 values of $F(2,24)$ are calculated. The only variable that yielded an F -to-enter exceeding the criterion was flexibility of closure, a high-level cognitive variable ($F(2,24) = 14.07$). This was somewhat higher than the F ratio this variable yielded before far vision was entered in the equation. All other

a subset of size 2 selected from m candidates by forward stepwise selection, the implied number of predictor equations examined is $m \times (m - 1)$ and the Bonferroni correction is $p \times m \times (m - 1)$. For $l = 3$, the implied number is $m \times (m - 1) \times (m - 2)$, and so on. Note that, if m is large and the predictors are correlated, this correction procedure rapidly becomes conservative as l increases.

³The F ratio for entering a variable into the prediction equation was the F from a one-way analysis of variance calculated using the variable's residuals, which is equivalent to the F produced by an analysis of covariance in which variables already in the prediction equation serve as covariates (Dixon et al., 1988).

variables had much smaller F -ratios (below 6.0). After flexibility of closure had been entered into the prediction equation, the values of $F(2,23)$ -to-enter for the remaining variables were substantially less than the next criterion value of 9.28 (3.15 and below).

To further investigate the variables distinguishing the three job groups, DAs were performed for each of the three possible pairwise contrasts between clusters. A criterion value of $F(1,21) = 17.875$ was adopted, which corresponded to the Bonferroni-corrected critical value of F that yields an experimentwise significance level of $p \leq .05$ when $3 \times 44 = 132$ F tests with 1 and 121 degrees of freedom are performed. The contrast between the logic-demanding technical jobs and pilot-like jobs yielded four skills with F ratios exceeding the criterion. These were far vision, glare sensitivity, and rate control. Spatial orientation was only slightly below criterion, with an $F = 16.42$. All of these skills were rated as more important to the pilot-like jobs. The contrast between pilot-like jobs and mechanical-technical jobs yielded no variables with F ratios exceeding the criterion. Flexibility of closure had the highest F ratio, 14.33. This variable was rated as more important for the pilot-like jobs than for the mechanical-technical jobs. Interestingly, the mechanical-technical jobs were scored more like pilot jobs than were logical jobs with respect to those skills that distinguished logical jobs from pilot jobs (except for depth perception). The contrast between the two technical job clusters also yielded no F ratios exceeding the criterion. The largest F ratio in this case was associated with mathematical reasoning ($F(1,21) = 12.42$), which was rated as more important for the logic-demanding jobs.

(U) Table VI. Clusters of Skills Obtained by Principal Components Analysis (Titles are Component Numbers in Order of Extraction). (U)

<u>PC1</u> Rate control Spatial orientation Glare sensitivity Far vision Depth perception Night vision Choice reaction time Visualization	<u>PC2</u> Trunk strength Dynamic strength Muscular flexibility Stamina Gross body coordination Sound localization	<u>PC3</u> Flexibility of closure Selective attention Speed of closure Perceptual speed Near vision
<u>PC4</u> Oral expression Oral comprehension Written expression Written comprehension	<u>PC5</u> Manual dexterity Finger dexterity Static strength	<u>PC6</u> Originality Problem sensitivity Mathematical reasoning
<u>PC7</u> Time sharing Number facility Written expression	<u>PC8</u> Color vision Night vision Reaction time	<u>PC9</u> Arm-hand steadiness Equilibrium
<u>PC10</u> Information ordering	<u>PC11</u> Memory Multi-limb coordination	<u>PC12</u> Logical reasoning

A second PCA was carried out to examine clusters among the skills. This PCA yielded 12 PCs with

eigenvalues greater than one. A skree test disclosed no obvious breakpoint in the plot of eigenvalue versus component number (see Fig. 2). The clusters defined by the skills' correlations with the 12 PCs ($r \geq 0.5$) are listed in Table VI. The first skill cluster contains several variables that were rated, on average, more important for the pilot-like jobs than for either of the technical jobs. The second skill cluster is dominated by a group of physical strength variables. These skills, on average, were rated somewhat more important for the mechanical-technical jobs than for the logical-technical jobs, and more important for the logical-technical jobs than for the pilot-like jobs. None of these skills discriminated well among the job clusters in the previous DAs. The third cluster contains a group of cognitive and sensory variables, among them flexibility of closure, a potentially discriminating variable identified in a previous DA.

The skills in this cluster were, on average, rated as somewhat more important for pilot-like jobs than for the logical-technical jobs, and more important for the logical-technical jobs than for the mechanical-technical jobs. The fourth cluster contains oral and written communication variables, none of which discriminated among the jobs. The fifth cluster contains two dexterity variables and static strength, which did not produce evidence of potential discriminating power. The sixth cluster is a set of cognitive variables that somewhat resembles cluster 3. Beyond this point, the clusters become increasingly difficult to interpret, suggesting that they may be largely noise.

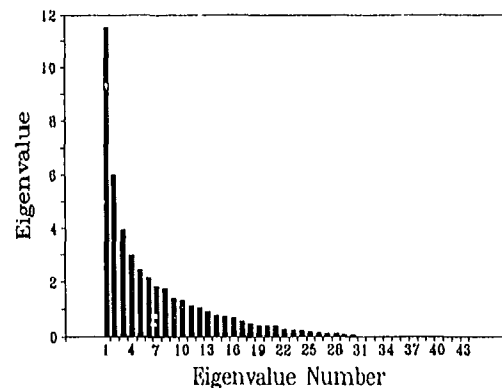


Figure 2. Eigenvalues from the principal components analysis of skills.

Six parallel discriminant analyses were performed on skill clusters identified by the PCA just described. The first was performed by forcing the apparently pilot-like skills of cluster 1 of Table VI into the equation, which yielded an approximate $F(10,42) = 5.038$. (This F ratio is an approximation to Wilks' λ that can be compared to ordinary F tables.) Had the pilot-like skills been selected by a priori criteria, the test would be significant at $p < .0006$, controlling, in Bonferroni fashion, for six, simultaneous F tests. A parallel discriminant analysis performed by forcing the strength-related skills of cluster 2 into the prediction equation yielded an approximate $F(12,40) = 1.32$. Even if the strength-related skills of cluster 2 had been selected by a priori criteria, this test would be nonsignificant. A third discriminant analysis performed by forcing the skills in cluster 3 into the prediction equation yielded an approximate $F(10,42) = 3.42$. Were the third cluster of skills selected by a priori criteria, this test would be significant at $p = .0138$, controlling for six tests. Discriminant analyses employing skill clusters 4 through 6 yielded F ratios of 3.02 and lower, which would also be nonsignificant.

Conclusions

The most important skills and abilities in the Cooper et al. (1987) data base of naval and marine combat occupations, as judged by their mean importance ratings, were a set of perceptual and cognitive abilities (see Table III). Coordination, dexterity, and orientation abilities tended to be rated as of intermediate importance. Strength and stamina variables tended to be rated as of lower importance. The skills and abilities that differed the most in importance from job to job were a group of perceptual, psychomotor, and strength skills, including far vision, rate control, stamina, glare sensitivity, and trunk strength.

Three primary clusters of combat occupations were tentatively identified in a principal components analysis (Table IV). The first cluster contained a set of logic-demanding technical jobs. The second contained pilot-like jobs. The third contained mechanical-technical jobs. The SEALs formed their own cluster. Two jobs were not assigned to any cluster. An exploratory analysis of the skills distinguishing these clusters suggests that the best discriminator could be far vision. Spatial orientation, rate control, and glare sensitivity may provide

lesser quantities of correlated predictive information. Somewhat surprisingly, the second best discriminator may be a cognitive skill, flexibility of closure. Discriminating between the pilot-like jobs and the logical-technical jobs was much easier than discriminating between the pilot-like jobs and the mechanical-technical jobs, or between the two clusters of technical jobs. Cognitive skills, as a group, displayed relatively little variation in rated importance from one job to the next.

Several clusters of skills and abilities were identified by principal components analysis (Table V). The first cluster was a group of apparently flight-related skills, including far vision, rate control, spatial orientation, and glare sensitivity. The second cluster contained a set of strength, stamina, and coordination skills, along with auditory localization. The third cluster of skills contained several perceptual and cognitive abilities, including flexibility of closure (a potentially discriminating ability) and selective attention (the ability with the highest overall importance rating). Unsurprisingly, given the preceding analysis, the first and third ability clusters gave some evidence of distinguishing between the job clusters; the remaining clusters did not.

One should bear in mind that abilities that discriminate between jobs need not be the most important abilities overall. Conversely, abilities that are important, overall, need not discriminate between jobs. Clearly, the most useful predictions of *differential* threat will occur in cases where a stressor is found to affect abilities that are uniformly important in some jobs and uniformly unimportant in others. An ability whose importance is unevenly distributed in this way is unlikely to be regarded as among the most important overall. In the present data, cognitive abilities were highly rated, as a group, yet the variability of their ratings across jobs was comparatively low (compare Tables III and IV).⁴ Thus, despite the uniformly high importance attributed to cognitive abilities, the present data suggest that abilities that were rated as of somewhat lower overall importance in these jobs might yield the best predictions of differential threats to performance.

3. SENSITIVITY OF TESTS

Discussion

Collins and Cliff (1990), in discussing the psychometric properties of growth measures, argue that the reliability coefficient, ρ_{XX} , is an inadequate measure of the precision of a test designed to assess change. They assert that the major traditions in psychometrics, classical test theory and item-response theory, have focused on the measurement of stable individual differences and have largely ignored issues surrounding the measurement of change. As a consequence, the major traditions have become "inadequate for and largely irrelevant to" the development of measures of growth (p. 128). Although strongly worded, this charge contains some truth. The association of ρ_{XX} with the ordinary-language terms "reliability" and "precision" has caused confusion and has occasionally led to inappropriate applications. The confusion stems from the idea that the reliability coefficient is generally applicable as a measure of the reliability and precision of a test. In fact, however, ρ_{XX} refers to a limited form of test reliability--specifically, the reliability with which a test detects stable differences among scores from different individuals.

The reliability coefficient is frequently characterized as an index of precision: Kerlinger (1986) refers to it as the "accuracy or precision of a measuring instrument" (p. 405). Lord and Novick (1968) call it the "imprecision and precision of tests" (p. 61). Introductory texts almost universally describe ρ_{XX} in similar terms. Common sense suggests that, if the reliability coefficient measures precision, a respectable value of ρ_{XX} is necessary for a test to be sensitive to change in a dependent variable. Thus, in discussing threats to statistical conclusion validity, Cook and Campbell (1979) assert that:

⁴The degree to which this may have been due to a compression of ratings at the upper end of the importance scale is an open question that warrants further attention.

Measures of low reliability (conceptualized either as "stability" or "test-retest") cannot be depended upon to register true changes. This is because unreliability inflates standard errors of estimates and these standard errors play a crucial role in inferring differences between statistics, such as the means of different treatment groups (p. 43).

Similarly, in describing a test battery designed to measure the effects of environmental stressors, the NATO Aerospace Medical Panel Working Group 12 (1989) concludes that:

Any psychological test must exhibit the properties of [construct] validity, reliability, and sensitivity. In other words, it must measure what it purports to measure, do so consistently, and be capable of detecting the effects of the environment or of individual differences in ability High [test-retest] reliability is a necessary, but not sufficient, condition for high validity. In other words, the target attribute cannot be measured adequately by a test that fails to provide consistent scores. . . (p. 6).

The idea that reliability is not sufficient to ensure validity is correct, as is the idea that a degree of "consistency" is necessary in a good test. However, the conclusion that a high reliability coefficient is necessary because consistent measurement is necessary is correct only when the purpose of testing is to measure individual differences. This is because a high value of ρ_{XX} means only that differences among individuals are large relative to measurement error: A high value of ρ_{XX} does not mean, however, that a test will consistently measure the effects of change in an experimentally manipulated independent variable.

This issue was addressed some years ago in a contentious and sometimes confusing interchange that began when Overall and Woodward (1975) offered the "paradoxical" observation that, when measurement error is held constant, the statistical power of a repeated measures analysis of change scores is maximized when the reliability coefficient of the scores is zero. To understand why this is so, recall that ρ_{XX} is traditionally defined as the proportion of the variance in a population of test scores attributable to variance in true scores (e.g., Gulliksen, 1950). That is:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{(\sigma_T^2 + \sigma_E^2)}, \quad (1)$$

where σ_T^2 represents the variance of the true scores, σ_X^2 is the variance of test scores, and σ_E^2 is the variance of the measurement errors. This definition is based on the assumption that each test score, X_i , is the sum of a true score, T_i , and measurement error, E_i (Gulliksen, 1950). The errors are usually assumed to be independent of the T_i , and of each other, and to have a mean of zero. Hence,

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \quad (2)$$

Perhaps the most widely used estimate of ρ_{XX} is the test-retest correlation, r_{X_1, X_2} . The test-retest correlation is estimated by obtaining test scores from the same group of individuals on two occasions and calculating the Pearson product-moment correlation between first and second scores.

Overall and Woodward (1975) considered the case of the t test for repeated (correlated) observations. When the null hypothesis is that the mean difference between scores obtained on two occasions is zero, the equation for the correlated t can be written:

$$t = \frac{\bar{d}}{\hat{\sigma}_d} \quad (3)$$

where \bar{d} is the mean of the differences and $\hat{\sigma}_d$ is the sample estimate of the standard error of the differences. Now, the variance of a set of differences is given by

$$\sigma_{\underline{d}}^2 = \sigma_{\underline{X1} - \underline{X2}}^2 = \sigma_{\underline{X1}}^2 + \sigma_{\underline{X2}}^2 - 2 \rho_{\underline{XX}} \sigma_{\underline{X1}} \sigma_{\underline{X2}}. \quad (4)$$

In Equation 4, $\rho_{\underline{XX}}$ is the correlation between pairs of scores in the two conditions and, hence, is equivalent to the population value of the test-retest correlation--that is, the reliability coefficient.

Assume, for the moment, that the variances of the measurement errors and true-scores are equal in the two experimental conditions. If so, then the variation attributable to individual differences in true scores disappears from the variance of the differences. To do this, set $\sigma_{\underline{X}}^2$ in Equation 4 to $\sigma_{\underline{T}}^2 + \sigma_{\underline{E}}^2$, which it equals by the definition expressed in Equation 2. This yields:

$$\begin{aligned} \sigma_{\underline{X1} - \underline{X2}}^2 &= 2\sigma_{\underline{T} + \underline{E}}^2 - 2\rho_{\underline{XX}} \sigma_{\underline{T} + \underline{E}}^2 \\ &= 2\sigma_{\underline{T} + \underline{E}}^2 (1 - \rho_{\underline{XX}}) \\ &= 2(\sigma_{\underline{T}}^2 + \sigma_{\underline{E}}^2)(1 - \rho_{\underline{XX}}) \end{aligned} \quad (5)$$

Now, solve Equation 1 for $\sigma_{\underline{T}}^2$ in terms of $\rho_{\underline{XX}}$ and $\sigma_{\underline{E}}^2$. The result is

$$\sigma_{\underline{T}}^2 = \rho_{\underline{XX}} \sigma_{\underline{E}}^2 / (1 - \rho_{\underline{XX}}). \quad (6)$$

Finally, replace $\sigma_{\underline{T}}^2$ in Equation 5 with the right-hand side of Equation 6. Simplifying produces the desired result:

$$\begin{aligned} \sigma_{\underline{X1} - \underline{X2}}^2 &= 2\{[\rho_{\underline{XX}} \sigma_{\underline{E}}^2 / (1 - \rho_{\underline{XX}})] + \sigma_{\underline{E}}^2\} (1 - \rho_{\underline{XX}}) \\ &= 2\{[\rho_{\underline{XX}} \sigma_{\underline{E}}^2 + (1 - \rho_{\underline{XX}})\sigma_{\underline{E}}^2] / (1 - \rho_{\underline{XX}})\} (1 - \rho_{\underline{XX}}) \\ &= 2[\sigma_{\underline{E}}^2(\rho_{\underline{XX}} - 1 - \rho_{\underline{XX}}) / (1 - \rho_{\underline{XX}})] / (1 - \rho_{\underline{XX}}) \\ &= 2\sigma_{\underline{E}}^2. \end{aligned} \quad (7)$$

Recall that the standard error of the differences is the square root of their variance divided by \underline{n} . Thus, given the null hypothesis $\mu_{\underline{X1} - \underline{X2}} = 0$, the \underline{t} for correlated observations can be written

$$\underline{t} = \frac{\underline{d}}{\sqrt{\text{Est.}[(2\sigma_{\underline{E}}^2) / \underline{n}]^{1/2}}}. \quad (8)$$

Equation 8 indicates that the sensitivity of a \underline{t} test for correlated observations (the magnitude of \underline{t}) depends only on the difference between the condition means (\underline{d}), sample size (\underline{n}), and the magnitude of measurement error ($\sigma_{\underline{E}}^2$). Neither the variance of the true scores nor the reliability coefficient of the test scores plays a role in the equation.

According to this analysis, subject-to-subject variation in difference scores is due only to measurement error. Were this always true, the reliabilities of difference scores would always be zero. This is because reliability, by the definition expressed in Equation 1, is the ratio of "true" variance to the sum of true and error variances, and because the "true" variance of difference scores is 0.0 under the model of Equation 7. Indeed, the reliabilities of empirical difference scores are frequently very low, a fact that has sometimes caused applied researchers to express concern about the wisdom of employing them in behavioral analyses.

Fleiss (1976) has argued, however, that to assume equal true and error variances in analyses such as the one just outlined is unrealistic. The reason is that this assumption ignores subject-to-subject variation in

responding to the independent variable. Fleiss noted, correctly, that scores in a one-way, repeated measures analysis of variance (ANOVA) frequently contain a third source of variation apart from treatment effects and measurement error. This variation arises from individual differences in responding to change in the independent variable and is the source of the subject-by-treatment interaction in a repeated measures ANOVA. The presence of this variation in empirical data is why the reliability coefficients of difference scores are not always zero: If a subject-by-treatment interaction exists, the variance of the differences will not be $2\sigma_E^2$, as Equation 7 asserts. Instead, the variance of the differences will equal $2(2\sigma_{SXT}^2 + \sigma_E^2)$, where σ_{SXT}^2 is the variance attributable to the subject-by-treatment interaction (Fleiss, 1976; Overall & Woodward, 1976; also see Winer, 1971, pp. 280-281). Hence the equation for \underline{t} when a subject-by-treatment interaction is present might be rewritten:

$$\underline{t} = \frac{\bar{d}}{\text{Est.}[2(2\sigma_{SXT}^2 + \sigma_E^2) / n]^{1/2}} \quad (9)$$

Fleiss (1976) outlined a repeated measures ANOVA model with a subject-by-treatment interaction and submitted that, in this model, when individual differences in responding are held constant, power is maximized not when the reliability coefficient of the scores is 0.0, but when reliability is 1.0. Note that the quantity $2\sigma_{SXT}^2 / (2\sigma_{SXT}^2 + \sigma_E^2)$ is the proportion of the variance of the difference scores that can be attributed to the subject-by-treatment interaction. By analogy with the definition expressed in Equation 1, this proportion can be understood to be the reliability of the difference scores, ρ_{dd} (Fleiss, 1976; Overall & Woodward, 1976).

Thus, according to Fleiss's (1976) analysis, when σ_{SXT}^2 is held constant and σ_E^2 is allowed to vary, the sensitivity of a \underline{t} or \underline{F} test calculated on the difference scores will necessarily vary directly with the reliability of the differences. This conclusion was consistent with that of Cleary and Linn (1969) and Sutcliffe (1958), who also varied error variance with true-score variance held constant, and concluded that the power of a significance test increases with the reliability of the dependent variable. Overall and Woodward (1976) replied to Fleiss's criticism by noting that the presence or absence of a term for σ_{SXT}^2 in the denominator of the \underline{t} ratio is irrelevant to the point they had originally made, which was that reliability is inversely proportional to the sensitivity of a \underline{t} test when measurement error is held constant. Adding a constant value (corresponding to the true variance of the difference scores) to the denominator of the \underline{t} ratio does nothing to change this basic algebraic relation.

A somewhat different result holds for between-subjects experimental designs. Suppose that an investigator is interested in determining whether an intervention affects true scores. For example, the investigator may be interested in determining whether a drug affects performance on some test. Assume that one group of subjects has been administered a placebo and the other has been administered the drug. The relation between reliability, error variance, and sensitivity for this contrast between group means is readily shown for the case of the \underline{t} test. When the null hypothesis is $\mu_1 - \mu_2 = 0$, the equation for an independent-samples \underline{t} test can be written

$$\underline{t} = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} \quad (10)$$

where $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$ is the sample estimate of the standard error of the difference between \bar{X}_1 and \bar{X}_2 . To keep the algebra simple, assume that the two groups are of equal size and that the true-score and error variances of the two groups are equal. Given these assumptions, the standard error of the difference can be rewritten as follows:

$$\begin{aligned} \sigma_{\bar{X}_1 - \bar{X}_2} &= [(\sigma_{X1}^2 + \sigma_{X2}^2) / n]^{1/2} \\ &= [(\sigma_{T1}^2 + \sigma_{E1}^2 + \sigma_{T2}^2 + \sigma_{E2}^2) / n]^{1/2} \end{aligned}$$

$$= [2(\sigma_T^2 + \sigma_E^2) / n]^{1/2} \quad (11)$$

Note that, in Equation 11, the "measurement error" obtained by decomposing the within-cell variance into σ_T^2 and σ_E^2 is not the same as the "experimental error" (MS_{error}) obtained by pooling within-cell variances in a conventional, one-way, between-subjects ANOVA (for an example, see Winer, 1971, p. 168). The ANOVA MS_{error} , like the variance in the denominator of the t test for independent samples, is derived from σ_X^2 . This quantity, under the traditional assumption of Equation 2, is the sum of σ_T^2 and σ_E^2 .

Replacing σ_T^2 in Equation 11 with the right side of Equation 6 allows us to rewrite the independent-samples t as follows:

$$\begin{aligned} t &= \frac{\bar{X}_1 - \bar{X}_2}{\text{Est.}\{[2(\rho_{XX} \sigma_E^2 / (1 - \rho_{XX}) + \sigma_E^2 / n)]^{1/2}\}} \\ &= \frac{\bar{X}_1 - \bar{X}_2}{\text{Est.}\{[2\sigma_E^2 / n(1 - \rho_{XX})]^{1/2}\}} \\ &= \text{Est.}\{[n(1 - \rho_{XX}) / 2\sigma_E^2]^{1/2}\}(\bar{X}_1 - \bar{X}_2), \end{aligned} \quad (12)$$

where $\text{Est.}\{\}$ refers to the sample estimate of the quantity inside the braces.

Equation 12 indicates that the value of t in the test for independent samples varies directly with the square root of $1 - \rho_{XX}$. Hence, for a constant σ_E^2 , the sensitivity of the t test for independent samples is a monotonically decreasing function of test reliability. The reduction in sensitivity that accompanies an increase in ρ_{XX} , other variables held constant, can be understood if one realizes that an increase in ρ_{XX} is tantamount to an increase in σ_T^2 relative to σ_E^2 (recall Equation 1). Thus, an increase in σ_T^2 that is not offset by a reduction in σ_E^2 will increase the total variance of the observations. An increase in σ_X^2 will, in turn, reduce the value of t , thereby reducing the sensitivity of the contrast.

Nicewinder and Price (1978) have pointed out that the conclusion one reaches regarding the effects of a difference in reliability on statistical power will depend on which variables are held constant when reliability is varied. If measurement error is held constant, an increase in reliability can only occur if the variance of the true scores increases, due to the relation expressed in Equation 1. An increase in σ_T^2 will tend to reduce the power of a between-subjects t or F test. This is because, as was discussed previously, both σ_T^2 and σ_E^2 enter the error terms of these tests.

If the variance of the true scores is eliminated by calculating difference scores, an increase in σ_T^2 cannot affect the power of a repeated measures test calculated using the differences. If, however, subjects vary in their responses to the independent variable, the interaction variance, σ_{SX}^2 , will be added to the variance of the differences. Under this condition, an increase in the reliability of the difference scores, ρ_{dd} , will reduce power. In contrast, if the variance of the true scores is held constant, an increase in reliability can only occur if measurement error decreases (Equation 1). Because σ_E^2 is always a component of the error terms of F and t tests, any reduction in measurement error will necessarily increase statistical power. Hence, if σ_T^2 is held constant, any increase in ρ_{XX} will increase statistical power.

In a consideration of the results of Overall and Woodward (1975, 1976) and Nicewinder and Price (1978), Sutcliffe (1980) argued that these authors had traded on a confusion of the reliability coefficient with the

"reliability of measurements" and "reliability proper" (p. 509), a point vigorously denied by Nicewinder and Price (1983). In his paper, Sutcliffe concedes that "power and reliability jointly depend on true and error variance" (p. 510). He concurs with Fleiss (1976), who, like Sutcliffe (1958) and Cleary and Linn (1969), concluded that power increases with reliability when individual differences are held constant and measurement error is varied. In discussing the result obtained when measurement error is held constant and individual differences are varied, however, Sutcliffe draws the spectacular conclusion that the inverse relation between power and reliability noted by Overall and Woodward (1975, 1976) is "spurious" (p. 513) and merely the result of change in "the numerical value of the [reliability] coefficient" (p. 513, emphasis added). A careful reading of Sutcliffe's (1980) paper suggests that, in fact, the author understands and concurs with Overall and Woodward's (1975, 1976) conclusion. He does not, however, make this point clear. Indeed, the text of Sutcliffe (1980) is easily misinterpreted as supporting the conclusion that ρ_{XX} and power are, in general, directly related.

Conclusions

Equations 8 and 12 show that the sensitivity with which a test measures change in true scores is not, in general, an increasing function of ρ_{XX} . Furthermore, the relation between sensitivity and ρ_{XX} differs in repeated measures and between-subjects experimental designs. In a one-way, repeated measures design with measurement error held constant, the reliability coefficient of the test scores will be unrelated to test sensitivity, provided no subject-by-treatment interaction exists. If a subject-by-treatment interaction is present, however, differences between-subjects' scores in some or all of the experimental conditions will have nonzero reliabilities. The "true" variance in the difference scores that accounts for this reliability will enter the denominators of t ratios calculated to test the significance of differences between the condition means. Although variance attributable to individual differences that remain constant across experimental conditions will not enter main-effects F ratios in conventional, repeated measures ANOVAs (see, for example, Winer, 1971), variance attributable to a subject-by-treatment interaction will enter both the numerator and denominator of such an F ratio and, thus, reduce statistical power (Winer, 1971).

In a between-subjects design, with measurement error held constant, an increase in ρ_{XX} will be accompanied by a reduction in sensitivity (Equation 12). This is because an increase in ρ_{XX} in the absence of a reduction in measurement error implies an increase in variance attributable to individual differences (by the definition expressed in Equation 1). This increase in true-score variance adds to the within-cell variance that enters the denominators of t tests (Equation 11), thereby reducing the tests' sensitivities. The increase in within-cell variance will also depress the F ratio for the corresponding main effect in a between-subjects ANOVA because the same within-cell variance components enter both the numerator and denominator of the F ratio (see Winer, 1971).

Even in the case of between-subjects designs, however, knowledge of the reliabilities of two tests is not, by itself, helpful in predicting which test is more likely to be sensitive to the effects of an independent variable. In part, this is because the magnitudes of reliability coefficients depend only on the relative magnitudes of true-score and measurement-error variances, whereas the magnitudes of between-subjects t and F tests depend on the (summed) absolute magnitudes of true and error variances (Nicewinder and Price, 1978). Nicewinder and Price (1978) have taken this observation to mean that no specifiable relation exists between reliability and sensitivity. Although this conclusion sounds discouraging, knowing both a test's reliability and the variance of its scores allows one to directly estimate the test's true and measurement-error variances: By the definition of Equation 1, $\sigma_T^2 = \rho_{XX}(\sigma_X^2)$ and $\sigma_E^2 = \sigma_X^2 - \sigma_T^2$. Thus, given estimates of a test's true and error variances, which can be obtained from test-retest data, one can use expressions such as Equations 7 and 11 to estimate the error terms the test would yield in various experimental designs.

A condition in which sensitivity varies as a direct function of reliability occurs when a test is changed in a way that affects only its measurement error. For example, the measurement error of a test can sometimes be reduced by increasing the length of a test. If other factors are held constant, a reduction in measurement error will simultaneously increase reliability and sensitivity (see Equations 1, 8, & 12). Nicewinder and Price

(1983) suggest that this relation may be the source of the erroneous belief that increased reliability is invariably associated with increased sensitivity. (They also provide a numerical example in which an increase in test length yields an increase in reliability accompanied by a reduction in sensitivity.) The existence of this special case, however, in no way implies that reliability and sensitivity to change in an independent variable will be positively associated in comparisons of arbitrarily selected tests. The sensitivity of a test to the effects of an independent variable is ultimately determined by the amount by which the test's scores change, relative to error, in response to change in an independent variable. The reliability coefficient, however, gives no indication of how much scores change in response to variation in an independent variable. Therefore, a difference in ρ_{XX} , by itself, gives no indication of the relative sensitivities of two tests.

Alternatives exist to the reliability coefficient as a measure of precision and sensitivity. Perhaps the simplest measure of precision is the standard deviation of the scores, σ_X . A limitation of σ_X , from the perspective of traditional test theory, is that it confounds variation due to individual differences with variation due to measurement error (see Equation 2). A better candidate might be the variance of the measurement errors, σ_E^2 (Dudek, 1979). As mentioned previously, σ_E^2 can be estimated from the same test-retest data used to estimate ρ_{XX} . When the absolute magnitudes of differences between the means of experimental conditions can be expressed in units that are comparable across tests, confidence intervals for the mean (or mean difference) can be useful measures of the precision and the potential sensitivity of a test. For example, the upper and lower limits of a $100(1 - \alpha)\%$ confidence interval for the difference calculated from test-retest data can be used to estimate the smallest changes in true scores that would be significant at a level of α in a within-subjects experiment with no subjects-by-treatment interaction.

A fundamental limitation that σ_E^2 shares with ρ_{XX} is that neither quantity reflects the degree to which a test's scores are likely to change in response to an intervention: A test that yields precise measurements of scores that do not change in response to an intervention will be less sensitive than a test that yields only rough measurements of scores that change substantially in response to the intervention. This means that, if you want to know which of several tests will most sensitive to the effects of an intervention, there is no substitute for pilot data--direct measurements of the effects of the intervention made with each test under consideration.

Signal detection theory provides several measures that can be used to compare the sensitivities of tests to the effects of change in an independent variable. The most familiar of these is the distance measure, d' , which is usually calculated as an estimate of $(\mu_2 - \mu_1) / \sigma_1$ (Green & Swets, 1966; Peterson & Birdsall, 1953). Cohen's (1977) measure of effect size, $d = (\mu_2 - \mu_1) / \sigma$, in which σ represents the standard deviation of the population corresponding to either μ_2 or μ_1 , is equivalent to d' when the populations' standard deviations are equal. The obvious repeated measures analog of these sensitivity measures is $\underline{d} = (\mu_2 - \mu_1) / \sigma_{X_2-X_1}$, in which the denominator is the standard deviation of a population of differences between paired observations. A straightforward generalization of \underline{d} to ANOVA is the ratio of treatment means to the common standard deviation, $\underline{f} = \sigma_\mu / \sigma$ (Cohen, 1977). The \underline{f} statistic is related ϕ , an index of effect size used in standard treatments of experimental power (e.g., Winer, 1971), by the relation $\underline{f} = \phi / n^{1/2}$ (Cohen, 1977). It is related to ω^2 , a widely used measure of ANOVA effect size, by the relation $\underline{f} = [\omega^2 / (1 - \omega^2)]^{1/2}$ (Keppel, 1991). The disadvantage of these measures, relative to the reliability coefficient, is that their calculation requires knowledge of an independent variable's effects on treatment means (to form estimates of $\mu_2 - \mu_1$ or σ_μ). This may not be a needlessly onerous requirement for a statistic that would be described as a measure of experimental "sensitivity."

Some of the persistence of the misunderstanding of the reliability coefficient may be attributable to differences in the perspectives of experimental and individual-differences researchers. Researchers into individual differences are accustomed to considering the effects of changes in measurement error on test results under the assumption that true-score variance remains constant. One must do this, for example, when estimating the effects of a change in test length. However, individual-differences researchers assuredly do not customarily think of differences among people as noise to be controlled and minimized. In contrast, experimental researchers frequently attempt to reduce the "noise" in their data by controlling the subject-to-subject variation in their

samples. (Recall from the preceding discussion that the "noise" variance in experimental data is frequently the sum of true-score and measurement-error variances.) This interdisciplinary difference in perspective may account for Fleiss's (1976) apparent failure to recognize Overall and Woodward's (1975) basic point (which was that the increase in true-score variance that accompanies an increase in reliability, when measurement error is held constant, reduces the sensitivities of experiments). A similar difference in perspective may account for Sutcliffe's (1980) assertion that the conclusions of Overall and Woodward (1975) and Nicewinder and Price (1978) might be taken to imply that "there may be an advantage in having noisy data" (p. 509). From the perspective of individual-differences research, a test that yields individual differences that are small, relative to measurement error, is a noisy test. This is true no matter how small measurement error is in absolute terms. From the perspective of experimental research, however, individual differences are likely to be regarded as noise whenever they inflate the error term of a significance test.

When individual differences are of interest, magnitudes of reliability coefficients are often measured and compared to justify the use of a particular test (Weiss and Davison, 1981). Indeed, Weiss and Davison (1981) refer to the measurement and comparison of reliability coefficients as a "preoccupation in psychometrics" (p. 633). They note that no other science has developed the concept of the reliability coefficient: all other disciplines express precision in terms of the probable error in measuring some true value. This meaning of "precision" is more nearly captured by the standard error of measurement, σ_E , than by the reliability coefficient. When change caused by variation in an independent variable is of primary research interest (and the measurement of individual differences is of lesser interest), comparing reliability coefficients to justify the use of a particular test is inappropriate and potentially misleading. This is because ρ_{XX} measures only the magnitude of the variation attributable to individual differences relative to the total variation in a data set. Many experimental studies are less concerned with differences among individuals than with changes within individuals. Examples include research into the effects of control and display configurations, training regimes, and environmental stressors. That a test reliably detects differences between individuals does not mean that it will necessarily perform "reliably" if it is turned to the measurement of within-person change, nor does a low value of ρ_{XX} imply that a test will be insensitive to change caused by an experimentally manipulated independent variable.

4. CARRIER LANDING MODEL

The line of research described in this section was interrupted by fiscal events that caused the project to be cancelled without notice. For that reason, the model is incomplete. The status of the model is described here at the Army's request. Our objective was to develop a model of cognitive workload in a carrier landing scenario. The immediate aims of this subproject were (a) to obtain a precise description of the time-course of human performance in what may be the most difficult aviation-related task and (b) to produce a quantitative description of moment-by-moment fluctuations in the workload imposed by this task. An ultimate purpose of this work was to identify variables in the carrier-landing scenario that may prove especially valuable in the development of valid and efficient designs for laboratory and flight-simulator research into medical issues in aviation performance.

Model Description

The information used to develop the aircraft carrier landing model was gathered via interviews with three pilot trainees at Pensacola Naval Air Station; therefore, a few tasks of the model are specific to training landings in T-2 aircraft. The model's overall network, called ACL2, is composed of 39 tasks and seven subnetworks. Each subnetwork comprises a set of tasks. The total number of tasks is 88. In Figure 1, a diagram of the overall network, tasks are represented by circles and subnetworks are represented by rectangles. Subnetworks of ACL2 in Figure 1 are numbered 22, 24, 23, 29, 30, 32, and 38.

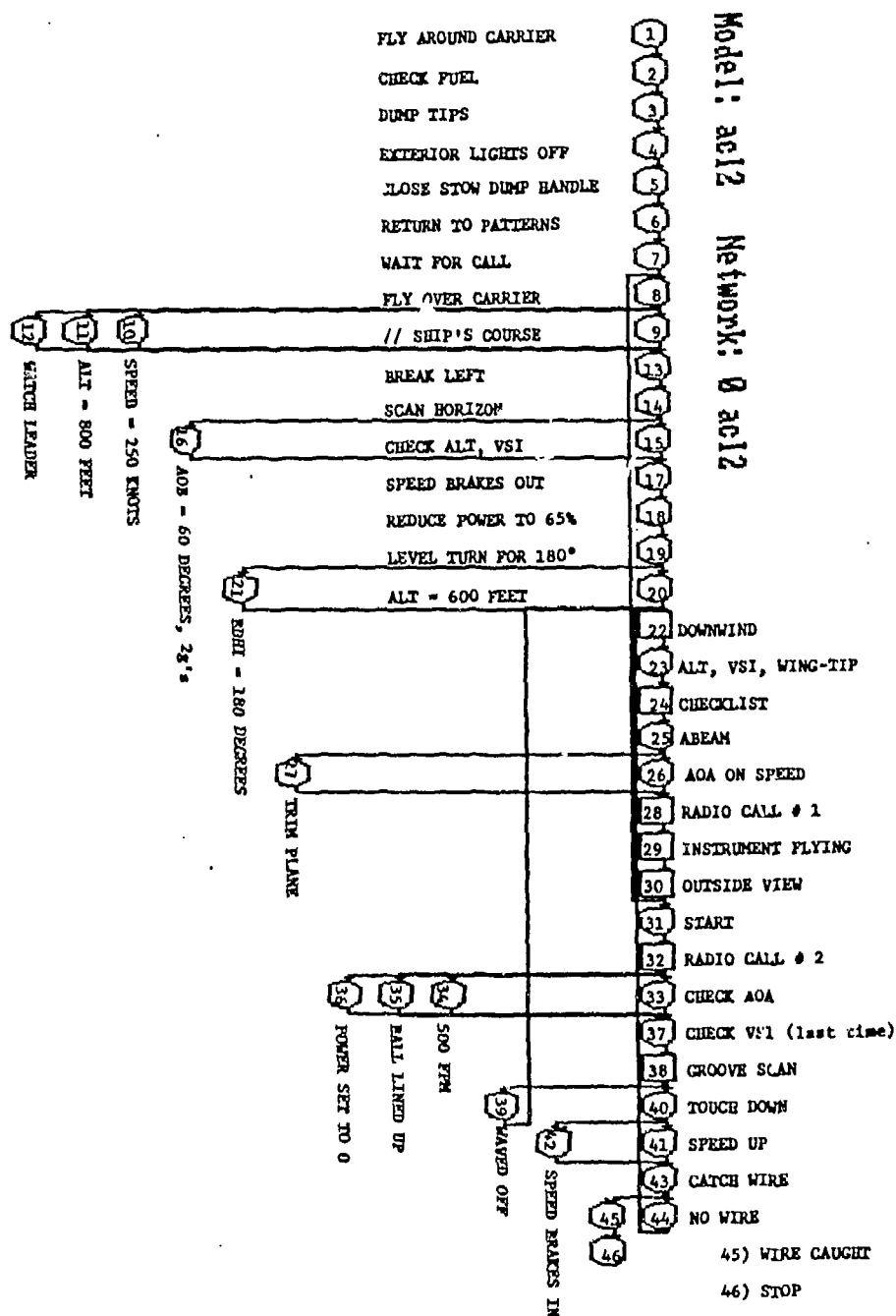


Figure 1. Overall Network Model

The first seven tasks are maneuvers completed by the T-2 pilots during training. The subsequent tasks are generally relevant to all aircraft carrier landings. The landing procedure begins when the pilot flies over the carrier. When the pilot has done this, he or she turns and parallels the ship's course, moving in the opposite direction, maintaining a speed of 250 knots and an altitude of 800 feet. Task 12 (Watch Leader) pertains to the training patterns of the T-2 pilots. Upon completing these tasks, the pilot breaks to the left and scans the horizon. He or she then attains an Angle of Bank (AOB) of 60 degrees at 2 Gs and checks his or her altitude and vertical slope index. If all is correct, the pilot extends the speed brakes (Task 17). Once the brakes are extended, the pilot reduces the power to 65%, and begins leveling the aircraft to 180 degrees from the landing point. Next, the pilot descends to 600 feet and obtains his or her position at 180 degrees from the landing point (The 180-degree measurement is relative to the direction of the carrier.)

When the pilot is flying at 180 degrees from the landing point, he or she begins "Downwind" procedures (subnetwork 22). This set of procedures comprises the first subnetwork of the model and consists of seven tasks. The layout of this subnetwork can be seen in Figure 2. Simultaneously, the pilot begins tasks 22.2 through 22.5. He or she achieves 15 units Angle of Attack (AOA--by varying speed, the critical value of which depends on aircraft weight, which varies with the amount of fuel remaining). He or she reduces air speed to 165 knots, prepares to get into the proper position abeam, lowers the flaps, and lowers the hook and the landing gear. This ends the first subnetwork.

The pilot then completes task 23 by checking altitude, the vertical slope index, and wing tip distance to confirm that all readings are correct. When these tasks are completed, the pilot performs a landing checklist. This checklist is the second subnetwork in the model. It comprises eight tasks (See Figure 3). Each checklist item constitutes a task. The pilot checks the amount of fuel in the aircraft, that the gear is down, that the flaps are down, that the hook is down, the harness is locked, and that the boards are working. He or she pumps the brakes and confirms that he or she is maintaining 15 units AOA. This process runs off very quickly. Upon completing this subnetwork, the pilot should be abeam. If so, the pilot checks to make sure the 15 units AOA setting is correct and simultaneously trims the plane.

If all readings are correct, the pilot enters the third subnetwork. Subnetwork 28 consists of the pilot's first radio call to the ship. The pilot must report the following: his or her side number, "gear, flaps" (meaning that these items have been checked and are down), his or her 15 units AOA at specific knots, his or her name, and his or her qualification number. Each call is a task, creating a total of seven tasks for this subnetwork (See Figure 4).

Following the radio call, the pilot enters the next subnetwork (See Figure 5). This subnetwork is called 'Instrument Flying' because the pilot performs the tasks almost entirely on instruments. The first three tasks (29.2, 29.3, and 29.4) of this network are performed simultaneously. The pilot checks that he or she is maintaining a proper AOA, that his or her altitude is at 600 feet, and that he or she is descending at 22 to 25 degrees AOB. At this point, the pilot begins a left-hand turn, placing him/herself at about 90 degrees from the landing point. The pilot descends to 450 feet, maintains his or her descent at 22 to 25 degrees AOB, and looks at the ship in order to correct any over- or under-shoot. In all, 10 tasks comprise this subnetwork.

The pilot then exits this subnetwork and enters another one called 'Outside View' (See Figure 6). During the set of six tasks that make up subnetwork 30, the pilot flies the aircraft using the outside view as a guide during about 70-80% of the flight time. When the pilot switches from instrument flying to the outside view, he or she should be at approximately 45 degrees from the landing point. The pilot checks and adjusts his or her AOB and references the "ball" on the carrier, which informs the pilot if he is coming in too high or too low. If the pilot does not see the ball, he or she must exit to task number 8, which entails flying over the carrier. If the pilot sees the ball, he or she decreases altitude to 275 to 300 feet.

At this point the pilot leaves the Outside View subnetwork and begins what the "Start" subnetwork, which comprises those tasks that take place on the final approach to the carrier. The pilot begins with a second

Model: ac12 Network: 22 Downwind

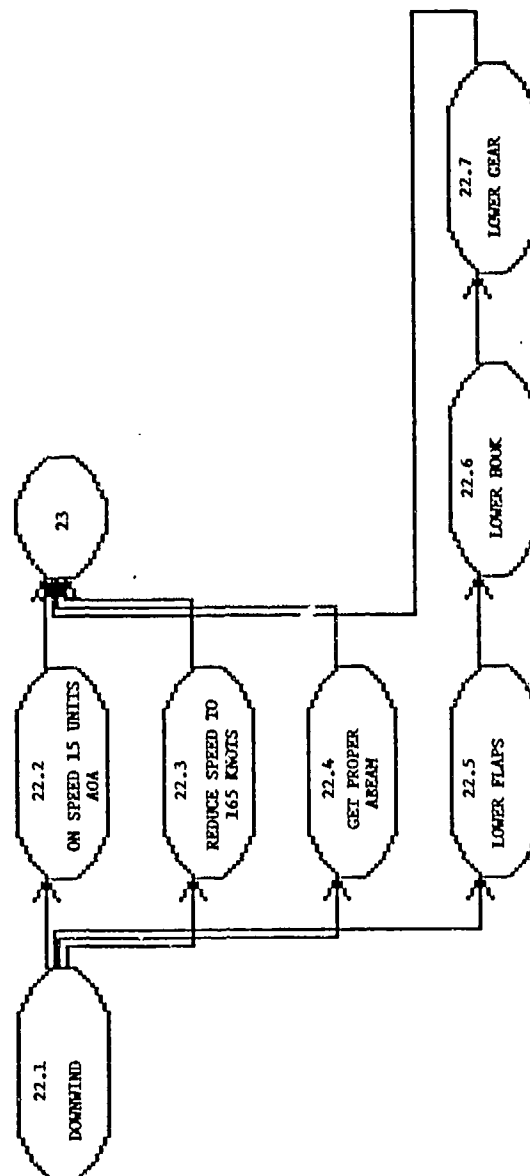


Figure 2. Subnetwork 22: Downwind Procedures

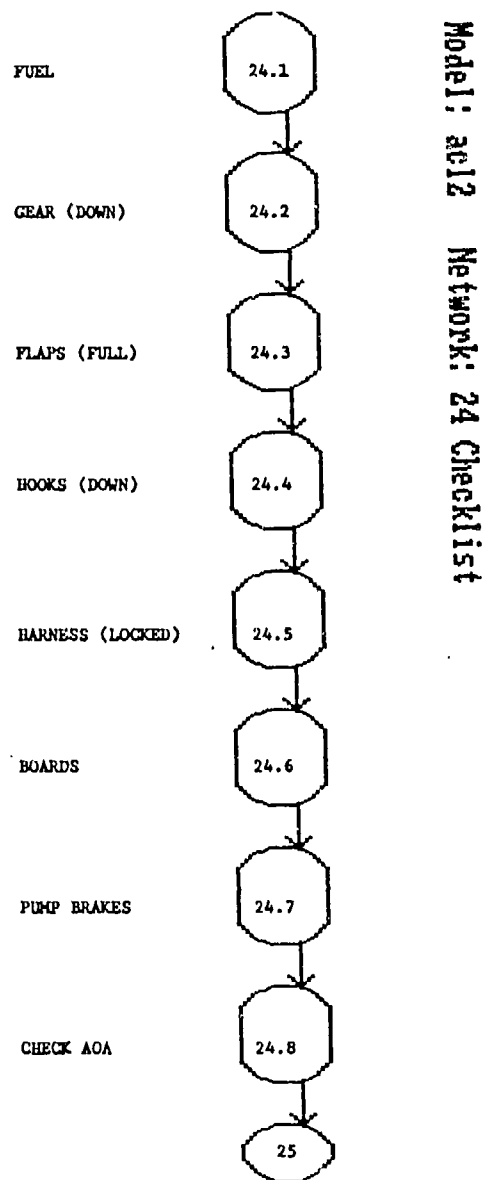


Figure 3. Subnetwork 24: Checklist # 1

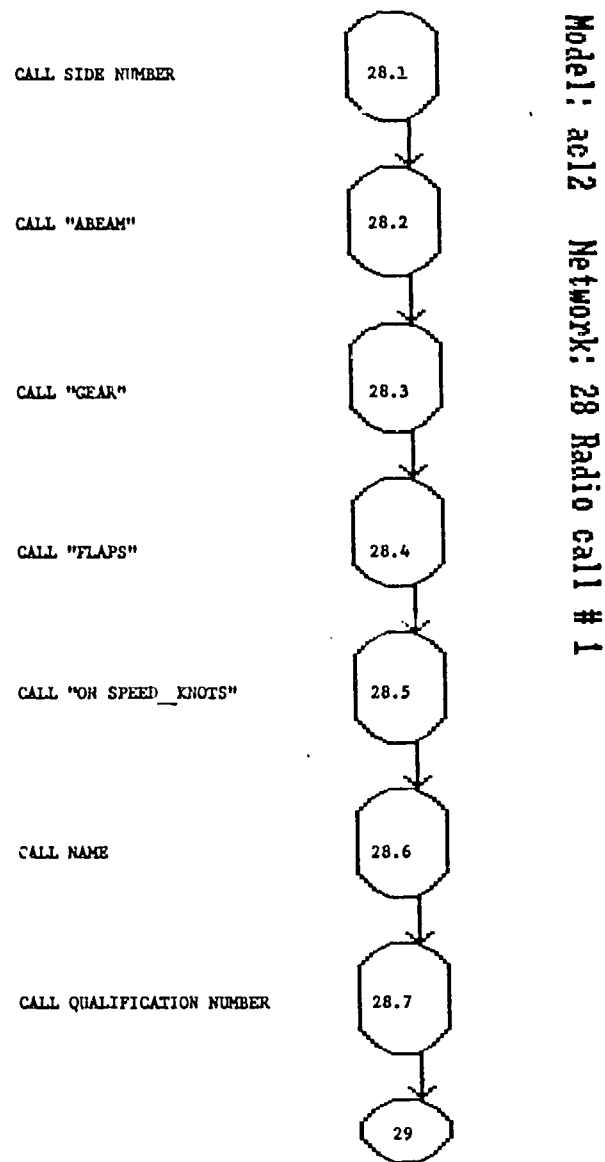


Figure 4. Subnetwork 28: Radio Call # 1

Model: ac12 Network: 29 Instrument Flying

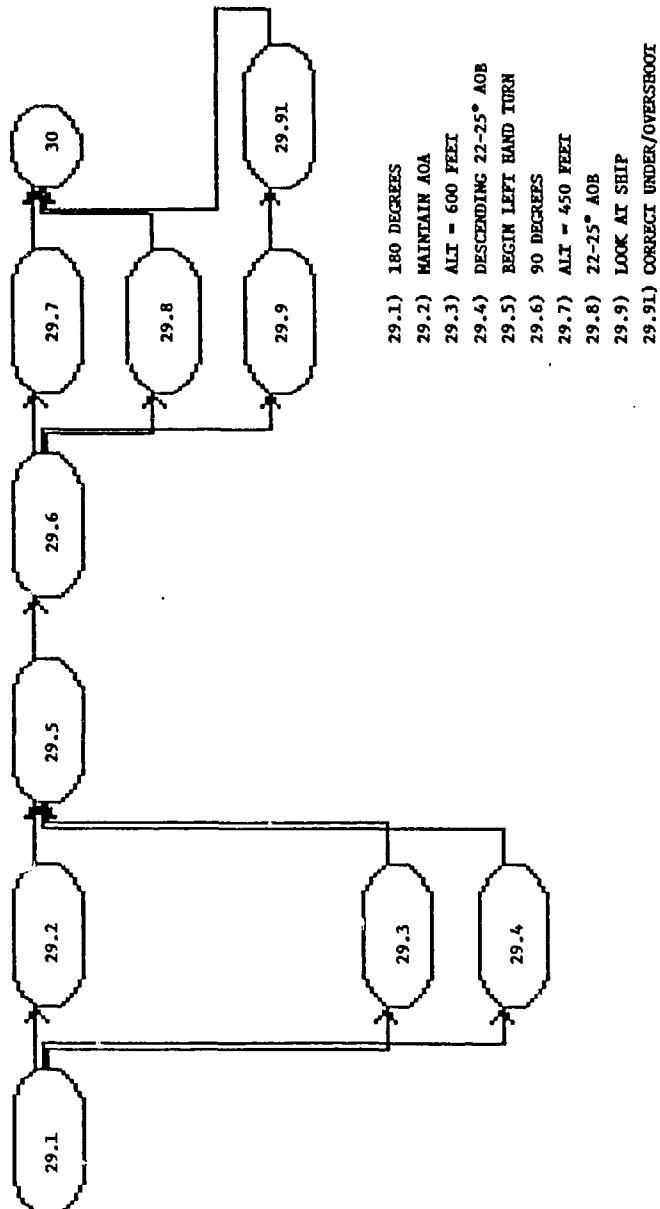


Figure 5. Subnetwork 29: Instrument Flying

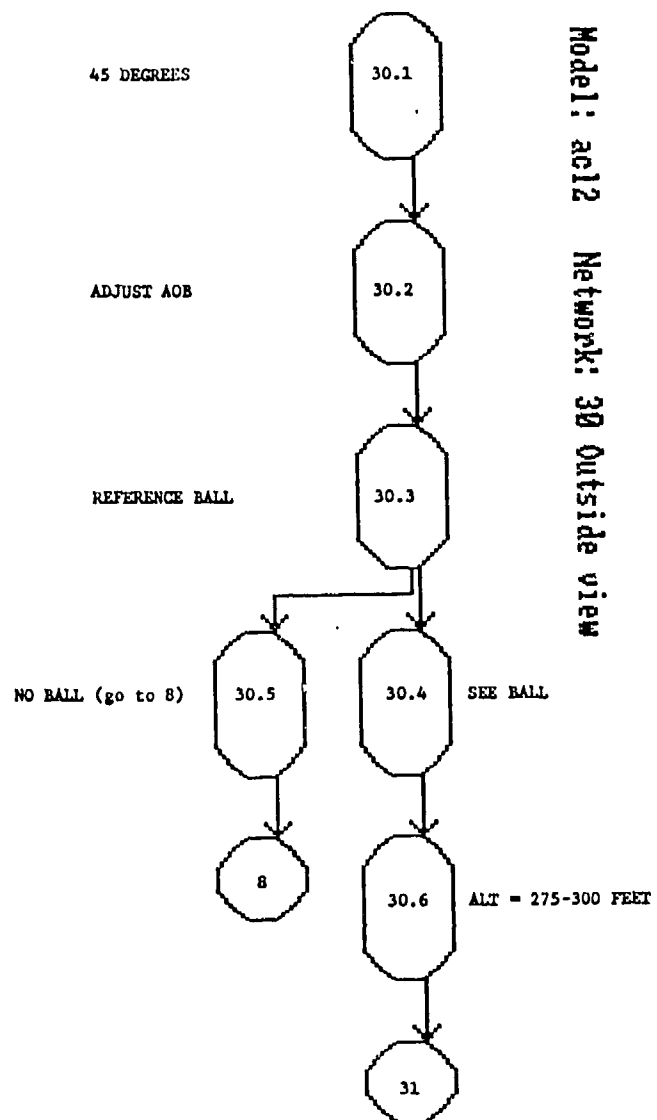


Figure 6. Subnetwork 30: Outside View Flying

radio call to the ship. This call is also represented by a subnetwork. The pilot gives his or her side number, squadron name, an indication that the ball has been seen, his or her fuel state, and his or her qualification number. All of these tasks must be completed in order to exit this subnetwork. (See Figure 7).

The pilot then engages in four simultaneous tasks. The pilot must check he or she is maintaining 15 units AOA (task 33), that he or she is descending at 500 feet per minute (task 34), that the ball is lined up (task 35), and that power is set to zero (task 36). Setting the power to zero does not mean the engines are turned off, but that the pilot is at appropriate throttle for the approach. Immediately following this, the pilot checks the vertical slope index for the last time.

The final subnetwork (# 38) of the model consists of several critical simultaneous procedures called the Groove Scan. The pilot repeatedly checks the ball, the AOA, and the line up, until the plane has landed. During this interval, the LSO informs the pilot by radio as to the accuracy of the approach and corrections that should be made. If the pilot fails to respond to an LSO instruction, the LSO waves the pilot off. In that event, the pilot must exit to the Downwind procedures (subnetwork 22) and attempt another landing from there. (For a diagram of these procedures, see Figure 8.) If the pilot responds appropriately, he or she exits the subnetwork and performs the final tasks, which take place on the carrier.

Tasks 40 through 42 consist of touching down on the carrier and immediately and simultaneously applying full thrust and pulling in the speed brakes. If the hook catches a wire, the pilot can shut down the aircraft. If the hook does not catch a wire, the pilot must perform a touch-and-go and return to the Downwind procedures (subnetwork 22) to try again.

5. BOOTSTRAP CONFIDENCE INTERVALS

Efron's bootstrap is a nonparametric technique for estimating variation in a statistic (Efron, 1978, 1982, 1988; Efron and Tibshirani, 1991). The procedure involves repeatedly drawing subsamples from an original data set. The statistic of interest is calculated in each subsample and the frequency distribution of its values is taken as an approximation to the statistic's actual sampling distribution.

The bootstrap is noted for wide applicability and a remarkable ability to extract information from samples (Efron, 1982). Several investigators have noted, however, that standard bootstrap confidence intervals (CIs) for the correlation coefficient yield overly liberal Type I error rates in small samples when α is set to .05 or less (e.g., Efron, 1982; Rasmussen, 1987, 1988; Strube, 1988). This bias may derive from a tendency of the bootstrap to produce too few subsamples with extreme values of the statistic under examination (Young & Daniels, 1990). Efron (1988) has observed that, although the bootstrap performs well with α set to .10, nonparametric bootstrap CIs perform better when "not pushed too far toward extreme coverage probabilities" (p. 295). In psychology and many other fields, however, it is conventional to set Type I error rates to .05 or less. Thus, Rasmussen (1987), Strube (1988), and others have advised caution in applying the bootstrap in small samples.

Bootstrap sampling is performed by randomly drawing observations from an empirical data set. Drawings are made with replacement and performed in such a way that each of the original observations has an equal probability of being drawn. The number of observations drawn for each subsample, n , is usually set equal to the number of observations in the original sample. The number of bootstrap subsamples drawn, N , varies with the problem. Nonparametric bootstrap CIs are typically based on 500-2000 subsamples; Efron (1988) suggests using a minimum of 1,000 subsamples.

A nonparametric, "percentile-method" bootstrap CI for an arbitrary statistic, θ , is generated by drawing N bootstrap subsamples, calculating the statistic's sample estimate, $\hat{\theta}$, in each subsample, finding the $100\alpha/2$ and

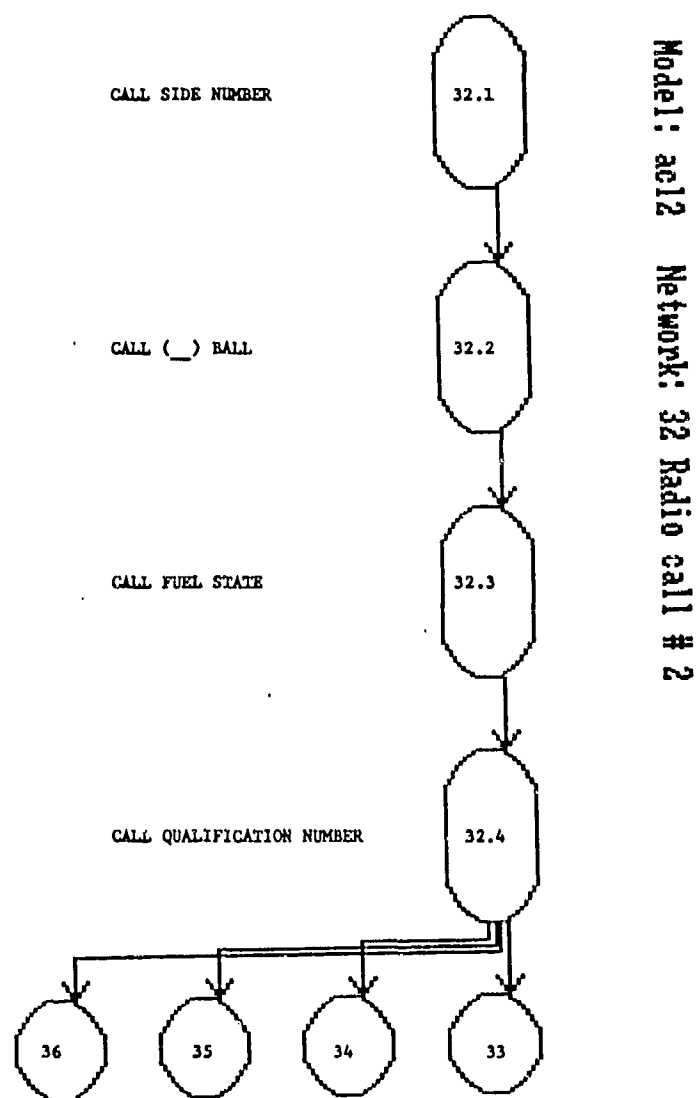


Figure 7. Subnetwork 32: Radio Call # 2

Model: ac12 Network: 38 Groove Scan

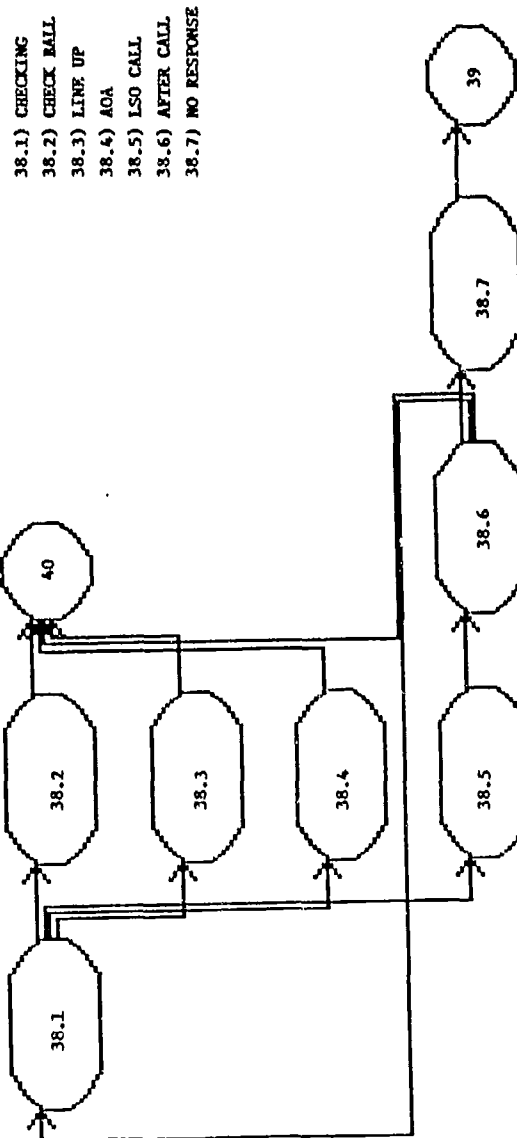


Figure 8. Subnetwork 38: Groove Scan

$100(1 - \alpha/2)$ percentiles of the frequency distribution of $\hat{\theta}$ -values thus produced, and taking the interval between these percentiles as the range of a $100(1 - \alpha)\%$ CI. In this way, the percentile method "automatically" determines approximate confidence limits associated with a given probability of Type I error. The data are not assumed to follow any specific probability distribution. The method does, however, depend on the assumption that distributions of bootstrap subsamples tend to reflect the forms of actual sampling distributions. The results of Rasmussen (1987, 1988) and Strube (1988) suggest that this assumption may be invalid when n is small and the desired α is .05 or less.

Several corrections for the percentile method's bias have been proposed. The bias-corrected percentile method (Efron, 1982) yields CIs with improved coverage properties. However, the bias-corrected percentile method reintroduces parametric assumptions. Furthermore, Monte Carlo studies have shown that the corrections it produces can be insufficient when n is small and α is set to .05 or less (Strube, 1988). The accelerated bias-corrected percentile method (Efron, 1987) can be quite accurate in some situations. This method has been criticized, however, for depending on an analytic correction factor that can be difficult or impossible to calculate (Loh & Wu, 1987).

Iterated bootstrapping is a computationally intensive approach to correcting the bootstrap's bias (Beran, 1987; Hall, 1986; Hall & Martin, 1988; Martin, 1990). Like the ordinary percentile method, the iterated percentile method sets confidence limits automatically, without parametric assumptions. An iterated bootstrap 95% CI for the mean can be calculated by drawing N first-order bootstrap subsamples from an empirical sample and M second-order subsamples from each first-order subsample. A percentile-method CI is derived from each of the N sets of second-order subsamples. The lower and upper cutoff percentages of the second-order CIs are adjusted until 95% of the intervals cover the sample mean. The adjusted cutoff percentages are then substituted for $100\alpha/2$ and $100(1 - \alpha/2)$ in an ordinary percentile-method CI for μ calculated from the means of the first-order bootstrap subsamples.

The Monte Carlo studies described here were performed to examine the Type I error rates of iterated-bootstrap 95% CIs for the mean in small samples drawn from non-normal populations. A CI for the mean expresses the precision with which a measurement has been obtained; it also can be used to test hypotheses about the location of μ . For example, a $100(1 - \alpha)\%$ CI for μ that does not cover 0 can be used to reject the null hypothesis that the data were sampled from a population with $\mu = 0$ at a significance level of α . Hence, the results of this study are directly relevant to one-sample hypothesis tests, such as tests of differences between correlated observations and 1-df orthogonal-polynomial contrasts.

Method

The simulations described here were written in Fortran-77 and run on an Intel 860 reduced instruction set processor installed in a desktop PC. Three types of CIs were examined: normal-theory (Student's t), percentile-method bootstrap, and iterated percentile-method bootstrap. Random samples of data were drawn from two distributions, one normal and one non-normal. Normally distributed samples were generated by drawing random values from a Gaussian distribution with $\mu = 0$ and $\sigma = 1$. Non-normal samples were generated by drawing random values from an exponential distribution with $\mu = \sigma = 1$. Gaussian random variables were generated by the direct method; exponential random variables were generated by the inverse method (Zelen & Severo, 1970).

Normal-theory and percentile-method CIs were compared in normally and exponentially distributed samples with sizes of 5, 10, 20, 40, 80, and 160. Iterated-bootstrap CIs were studied in normally and exponentially distributed samples with sizes of 5, 10, 15, and 20. One thousand confidence intervals were created in each experimental condition defined by a combination of CI type, probability distribution, and sample size. The observed Type I error rate in each experimental condition was calculated as the proportion of CIs that failed to cover μ .

A normal-theory CI was calculated as $\bar{x} \pm t_{(1-\alpha/2)}(s_{\bar{x}})$, where $t_{(1-\alpha/2)}$ is the critical value of Student's t corresponding to $1 - \alpha/2 = .975$, with $n - 1$ degrees of freedom, and $s_{\bar{x}}$ was the sample estimate of the standard error of the mean. A percentile-method CI was calculated by drawing $N = 1,000$ bootstrap subsamples from an empirical sample, calculating the mean of each subsample, and taking the 2.5th and 97.5th percentiles of the subsample means as the $100\alpha/2\%$ and $100(1 - \alpha/2)\%$ limits of the 95% CI, respectively.

An iterated-bootstrap CI was calculated by drawing $N = 1,000$ first-order bootstrap subsamples from an empirical sample, drawing $M = 1,000$ second-order bootstrap samples from each first-order sample, and finding the means of the second-order samples. The N cumulative frequency distributions of second-order subsample means were searched for the percentile that exceeded the sample mean in 2.5% of the distributions. A similar search was performed for the percentile that exceeded the sample mean in 97.5% of the distributions. An ordinary percentile-method CI was then constructed from the means of the first-order subsamples, with $100\alpha/2\%$ and $100(1 - \alpha/2)\%$ replaced by the percentages found in the search through the second-order means.

Results and Discussion

Figure 1 illustrates the performance of the normal-theory and percentile-method CIs in normally distributed samples. The Type I error rates of the normal-theory intervals are near the nominal α level of .05 at all sample sizes. In contrast, the Type I error rates of the percentile-method intervals are substantially higher than .05 in small samples, averaging about .153 when $n = 5$, and do not approach .05 until n reaches about 40. In samples of about 40 or more observations, the percentile-method bootstrap works quite well. Indeed, the average CI limits, not shown, are near those given by normal-distribution theory. In small samples, however, the percentile method yields limits that tend to be narrower than those given by theory, a result that accounts for the disproportionate numbers of Type I errors.

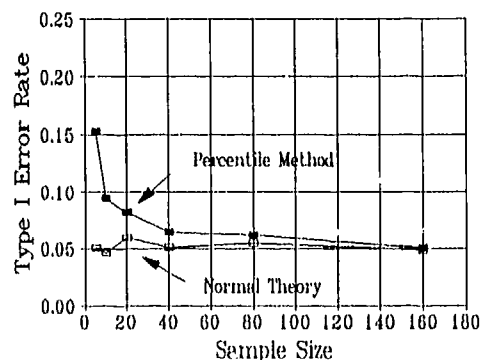


Figure 1. Empirical Type I error rate as a function of sample size for percentile-method bootstrap and normal-theory confidence intervals. The nominal protection level was 95%. Samples were drawn from a Gaussian distribution.

Figure 2 illustrates the performance of the normal and percentile-method bootstrap intervals in samples drawn from an exponential distribution. Both types of interval produce Type I errors at rates much larger than .05 in samples smaller than 20. Interestingly, the parametric intervals are less biased than the nonparametric intervals. Neither interval, however, performs especially well in the sample sizes examined here.

Figure 3 illustrates the performance of the iterated-bootstrap intervals in samples drawn from normal and exponential distributions. In samples of 5 observations, the iterated-bootstrap's Type I error rate is still high, averaging about .096 in normal samples and about .139 in exponential samples. In samples of size 10, however, Type I errors approach $\alpha = .05$ in normal samples (averaging .062) and are only slightly higher in exponential samples (averaging about .066 in the exponential data). In samples of 15 or more observations, the iterated bootstrap's observed Type I error rate is near $\alpha = .05$ in both normal and exponential samples. (The normal approximation to the binomial suggests that the standard error of the estimate of $\alpha = .05$ in samples of 1,000 should be about $[(\alpha)(1 - \alpha)/1000]^{1/2} = .007$.)

Conclusions

None of the methods yielded Type I error rates near $\alpha = .05$ in samples of 5 observations from an exponential distribution. Iterated-bootstrap CIs produced liberal Type I error rates in samples of fewer than about 10 observations. Normal-theory CIs were similarly biased in samples of fewer than about 20 observations. Ordinary percentile-method bootstrap CIs were seriously biased in samples of fewer than about 40 observations.

Except in the smallest samples ($n = 5$), the iterated-bootstrap intervals yielded Type I error rates in Gaussian data that were nearly indistinguishable from the Type I error rates of the normal-theory intervals. This is a remarkable performance for a nonparametric technique, given that the normal-theory CIs are optimal in Gaussian data. The failure of the iterated bootstrap in the $n = 5$ condition is disappointing but unsurprising, given the uncertainties involved in reconstructing the sampling distribution of the mean from so few observations.

An iterated bootstrap requires substantially more computer time than the ordinary bootstrap, which is itself computationally demanding. When N and M are set to 1,000, for example, an iterated bootstrap requires drawing $NM = 1,000,000$ subsamples, calculating 1,001,001 values of the statistic of interest, and sorting 1,001 arrays of 1,000 means. Some additional time is spent searching the arrays of second-order means for the adjusted percentile cutoffs. Calculations of this size take time, but are not beyond the capabilities of personal computers. A problem with 15 observations and values of $N = M = 1,000$ should take about 5 min on an Intel 80386/20-based computer with a math coprocessor, which may be less time than would be necessary to calculate the equivalent normal-theory CI with pencil and paper. Larger problems would be proportionally more time consuming. These, however, could be handled by the percentile method, or by normal-theory techniques.

Two practical recommendations for data analysis are suggested by the results. First, ordinary percentile-method bootstrap CIs for μ may be of questionable value when Type I error rates are to be controlled at values as low as .05. This is because, when $\alpha \leq .05$, percentile-method CIs may perform less well than parametric CIs in small samples and no better than parametric CIs in large samples. Second, when a data set may have been drawn from a skewed distribution, such as the exponential, iterated-bootstrap CIs may be preferable to parametric CIs if n is about 10 or more. Under these conditions, iterated CIs may yield better levels of Type I error control than parametric CIs when the data are skewed, and approximately the same Type I error control when the data are normal.

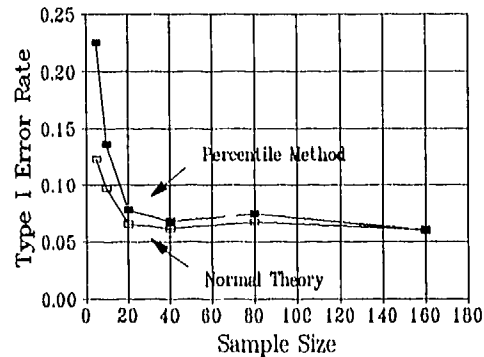


Figure 2. Empirical Type I error rate as a function of sample size for percentile-method bootstrap and normal-theory confidence intervals. The nominal protection level was 95%. Samples were drawn from an exponential distribution.

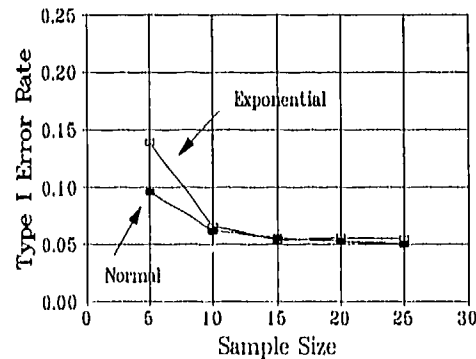


Figure 3. Empirical Type I error rate as a function of sample size for iterated percentile-method bootstrap confidence intervals. The nominal protection level was 95%. Samples were drawn from normal and exponential distributions.

REFERENCES

- Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, 74, 457-468.
- Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Statistical and Mathematical Psychology*, 22, 49-55.
- Cliff, N., (1987). *Analyzing Multivariate Data*. San Diego, CA: Harcourt Brace Jovanovich.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Orlando, FL: Academic Press, Inc.
- Collins, L. M., & Cliff, N. (1990). Using the longitudinal Guttman simplex as a basis for measuring growth. *Psychological Bulletin*, 108, 128-134.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally College Publishing Co.
- Cooper, M., Schemmer, M. S., Fleishman, E. A., Yarkin-Levin, K., Harding, F. D., & McNelis, J., (1987). *Task analysis of Navy and Marine Corps occupations: A taxonomic basis for evaluating CW antidote/pretreatment drugs*. (Report No. 3130). Bethesda, MD: Advanced Research Resources Organization.
- Dixon, W. J., Brown, M. B., Engelman, L., Hill, M. A., & Jennrich, R. I. (1988). *BMDP statistical software manual*. Berkeley: University of California Press.
- Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, 86, 335-337.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.
- Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82, 171-185.
- Efron, B. (1988). Bootstrap confidence intervals: Good or bad? *Psychological Bulletin*, 104, 293-296.
- Efron, B., & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science*, 253, 390-395.
- Fleishman, E. A., & Quaintance, M. K., (1984). *Taxonomies of human performance*. Orlando, FL: Academic Press.
- Fleiss, J. L. (1976). Comment on Overall and Woodward's asserted paradox concerning the measurement of change. *Psychological Bulletin*, 83, 774-775.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Huntington, NY: Robert E. Krieger Publishing Co.
- Guilford, J. P., (1954). *Psychometric Methods*. New York: McGraw-Hill.

- Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.
- Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics*, 14, 1431-1452.
- Hall, P., & Martin, M. (1988). On bootstrap resampling and iteration. *Biometrika*, 75, 661-671.
- Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Prentice Hall, Englewood Cliffs, NJ.
- Kerlinger, F. N. (1986). *Foundations of behavioral research*. New York: Holt Rinehart and Winston.
- Larzelere, R. E., & Muliak, S. A. (1977). Single-sample tests for many correlations. *Psychological Bulletin*, 84, 557-569.
- Loh, W.-Y., & Wu, C. F. J. (1987). Comment on "Bootstrap confidence intervals and bootstrap approximations." *Journal of the American Statistical Association*, 82, 188-190.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Martin, M. A. (1990). On bootstrap iteration for coverage correction in confidence intervals. *Journal of the American Statistical Association*, 85, 1105-1118.
- NATO Aerospace Medical Panel Working Group 12. (1989, May). *AMP Working Group 12 and AGARD lecture series 163, AGARDograph No. 308: Human performance assessment methods*. Neuilly-Sur-Seine, France: North Atlantic Treaty Organization Advisory Group for Aerospace Research and Development.
- Nicewinder, W. A., & Price, J. M. (1978). Dependent variable reliability and the power of significance tests. *Psychological Bulletin*, 85, 405-409.
- Nicewinder, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests: Some new results. *Psychological Bulletin*, 94, 524-533.
- Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, 82, 85-86.
- Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin*, 83, 776-777.
- Peterson, W. W., & Birdsall, T. G. (1953). *The theory of signal detectability*. University of Michigan: Electronic Defense Group, Technical Report No. 13.
- Rasmussen, J. L. (1987). Estimating correlation coefficients: Bootstrap and parametric approaches. *Psychological Bulletin*, 101, 136-139.
- Rassumssen, J. L. (1988). "Bootstrap confidence intervals: Good or Bad": Comments of Efron (1988) and Strube (1988) and further evaluation. *Psychological Bulletin*, 104, 297-299.
- Stanny, R. R. and Shamma, S. E., (1990). *Models of Human Performance Assessment Tests*. NAMRL Monograph, Naval Aerospace Medical Research Laboratory, Pensacola, FL. In review at the Naval Aerospace Medical Research Laboratory.

- Strube, M. J. (1988). Bootstrap Type I error rates for the correlation coefficient: An examination of alternate procedures. *Psychological Bulletin*, 104, 290-291.
- Subkoviak, M. J., & Levin, J. R. (1977). Fallibility of measurement and the power of a statistical test. *Journal of Educational Measurement*, 15, 111-116.
- Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika*, 23, 9-17.
- Sutcliffe, J. P. (1980). On the relationship of reliability to statistical power. *Psychological Bulletin*, 88, 509-515.
- Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology*, 32, 629-658.
- Winer, B. J. (1971). *Statistical principles in experimental design*, New York: McGraw-Hill.
- Young, G. A., & Daniels, H. E. (1990). Bootstrap bias. *Biometrika*, 77, 179-185.
- Zelen, M., & Severo, N. C. (1970). Probability functions. In M. Abramowitz and I. A. Stegun (Eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (pp. 925-995). Washington, DC: U.S. Government Printing Office.

BIBLIOGRAPHY

- Stanny, R. R. and Shamma, S.E., *Models of Human Performance Assessment Tests*, NAMRL Monograph, Naval Aerospace Medical Research Laboratory, Pensacola, FL. In review at the Naval Aerospace Medical Research Laboratory.
- Shamma, S. E., Stanny, R. R., and Morey, W.A., *Micro SAINT Modeling of Physiological Responses and Human Performance in the Heat*, NAMRL Monograph 42, Naval Aerospace Medical Research Laboratory, Pensacola, FL. Naval Aerospace Medical Research Laboratory, Pensacola, FL, March 1991.
- Stanny, R. R., *Reliability and Sensitivity Revisited*. In internal review at NAMRL.
- Stanny, R. R., *Predicting Stressor Effects in 31 Naval Combat Occupations*. In internal review at NAMRL.
- Stanny, R. R., *Bootstrap Type I Error Rates: Confidence Intervals for μ in Small Normal and Exponential Samples*. In Preparation.
- Shamma, S., and Stanny, R. *Models of Cognitive Performance Assessment Tests*, Presentation at the Eighth International Conference on Mathematical and Computer Modeling, University of Maryland at College Park, College Park, MD, 1-4 April 1991.

RELATED PUBLICATIONS

- Stanny, R., Shamma, S., Laughery, R., Platt, C., Crisman, R., and Sherry, D. "Modeling the Unified Tri-Service Cognitive Performance Assessment Battery." In *Proceedings of the 1989 Medical Defense Bioscience Review*, Columbia, MD, August, 1989, pp. 793-796.
- Shamma, S., Stanny, R., Laughery, R., Platt, C., and Sherry, D., *Computer Aided Modeling of Cognitive Performance Assessment Tests Using the Micro SAINT Software*, Institute for Statistical and Mathematical Modeling Technical Report, University of West Florida, Pensacola, March, 1989.
- Shamma, S. E., Molina, E. A., and Stanny, R. R., *Micro SAINT Programs for Numerical Methods of Integration and Differentiation*, NAMRL Monograph 39, Naval Aerospace Medical Research Laboratory, Pensacola, FL, September, 1989.